Data Science and Machine Learning Self-Guided Capstone Laboratory (D2K Learning Lab)

Data Science 400 • Spring 2020 • Christopher Tunnell

Time & Location:
Mondays & Wednesdays 5:00pm - 6:30pm
Location: Rayzor 113

Instructor:
Christopher Tunnell (tunnell@rice.edu)

D2K Fellows:
Binghan Yuan (by8@rice.edu)

D2K Lab Administrator:
Jennifer Sanders (Jennifer.Sanders@rice.edu or d2k@rice.edu)

Course Description:
In this project-based course, student teams will complete semester-long data science research or analysis projects based on self-chosen data sets. Students will also learn best practices in data science.

Course Overview:
In this project-based course, student teams will complete semester-long data science research or analysis projects based on self-chosen data sets. Each student team will have a Rice technical mentor (faculty and/or D2K Fellow). Student teams are required to meet weekly outside of class with their mentors. During the semester, teams will work towards project milestones and present their progress in interim reports, oral presentations, and software checks. At the end of the semester, each team will produce a final project report with accompanying software and give a poster presentation on their work in the D2K Showcase. This course will also have a lecture component at the beginning of the semester where students will learn how to design a data science pipeline and best practices in data science with emphasis placed on computational and data scientific reproducibility.

Course Objectives & Outcomes:
In this course, students will engage in experiential learning where they learn best practices in data science by working on real-world data science projects. Students will learn how to wrangle, explore, and model their data as well as how to design a reproducible data science pipeline and validate their data science findings. Through these projects, students will also learn team-work, leadership, project management, and communication skills. Students completing this course should be able to (i) successfully complete a data science research or analysis project, (ii) successfully devise and employ a computationally and scientifically reproducible data analysis pipeline, (iii) work collaboratively on a team to complete a data science project, write reports, and give presentations, and (iv) effectively communicate their findings to both a general and a technical audience.

DSCI 400 v.s. [(DSCI 435/535) (COMP 449/549)]:

DSCI 400 is a self-guided curiosity-driven project course. Guidance will be given on available datasets and project directions, but the teams will be responsible for defining a proposal, getting that proposal approved by the instructor, and then implementing it. This contracts the other parts of the D2K Learning Lab that are client facing with client-defined projects. One target audience of DSCI 400 is those completing the Data Science minor. As DSCI 400 is a 3 credit course, please consult your department to check that this course satisfies any degree requirements. Graduate students should not enroll in DSCI 400.

Course Application & Registration:
This course requires an application and instructor permission for registration. While students can pre-register for the course before the semester begins, registration will only be finalized after students have applied and been placed on a team during the first week of the semester. Typically, the first class (Monday at 5pm) will include an overview of the available projects and the course, with instructors and D2K Fellows available to answer questions. Then, a course application will be due by the next Tuesday at 12pm; in the application, students will be asked

to highlight their data science background and experience as well as list their project preferences and interests. Students will be assigned to teams (and their course section) the next Wednesday before the first official class.

Team & Project Placement:
Students will be selected for specific projects based on their background and skill set as well as their project preferences.

Lectures:
At the beginning of the semester, there will be a small number of lectures on the data science pipeline, how to design a pipeline, technology for data science, and data science reproducibility (see course schedule below).

Coursework:
The major work of this course will be to conduct a data science research or analysis project. Each team will present their work through poster presentations at the D2K Showcase and through a final report with accompanying software due at the end of the semester. For this course, there is no oral presentation at the D2K Showcase. Throughout the semester, shorter interim reports, presentations, and software checks will be due to give each team feedback and guidance on how best to prepare their final report and presentation. Further details and instructions on the reports, presentations, and software checks as well as grading rubrics are given in a separate document.

Course Expectations:
Students are expected to work at least 7 hours a week on their project outside of class. Each week outside of class, teams are required to have one meeting with their mentors and at least one separate meeting with just team members. The D2K administrator will help each team with scheduling meetings and booking conference rooms. Attendance at team meetings and lectures is required. Additionally, each team member is expected to contribute to all aspects of the data science project including technical data science contributions, coding, and preparing the written report and presentations.

Grading Overview:
Individual: 20%
- Peer evaluations: 10%
- Individual contributions: 10%

Team: 80%
- Project Definition: 10%
- Interim Evaluations: 15%
- Final Report: 25%
- Final Presentation: 15%
- Final Software: 15%

Grading rubrics as well as details on all coursework are given in a separate document.

Team Participation: Attendance and team participation will be evaluated through peer evaluation surveys at CATME.org. Students will need to register on this site to access the peer evaluations.

Computing Resources:
The D2K Lab has partnered with Amazon Web Services (AWS) to allot each client-facing team $500 of computing credits. Teams can request additional credits as needed.

Course Schedule:

Date	Class Time	Coursework Due
13-Jan	Welcome Pizza Party!	
15-Jan	Intro + Meet teams, and mentors!	
22-Jan	Lecture: DS Pipeline	Data Set and Project - Proposal Ideas Due
27-Jan	Lecture: Designing a DS Pipeline	
29-Jan	Lecture: Technology for DS Pipelines	Project Proposal Due
3-Feb	Lecture: DS Reproducibility	
5-Feb	Weekly Check-ins	
10-Feb	Lecture: Data Exploration	Initial Report (intro + background + objectives + data description + DS pipeline design plan)
12-Feb	Presentations (DS Pipeline Design Plans) + Weekly Check-ins	Initial Presentations (DS Pipeline Plan)
17-Feb	Lecture: Data Exploration	
19-Feb	Initial Presentations + Weekly Check-ins	Initial Presentations (Intro, Background, Objectives, DS Pipeline Plan, Data, Data Wrangling, beginning Data Exploration)
24-Feb	Weekly Check-ins	
2-Mar	In-Class software checks + Weekly Check-ins	Midtern Software Check
9-Mar	Midterm Presentations	Midterm Presentation + Midterm Report Due + Midterm CATME Eval Due
23-Mar	Weekly Check-ins	
30-Mar	Weekly Check-ins	Midterm Report Revision Due
6-Apr	Presentations + Weekly Check-ins	Interim Presentation
13-Apr	Software Checks + Weekly Check-ins	Interim Software Check
20-Apr	Final Presentations + Weekly Check-ins	Final Oral Presentation
22-Apr	D2K Showcase	Final Poster Presentation
6-May	End of Semester	Final Report + Final Software Due + Final CATME Eval Due

Students with a documented disability requiring accommodations should speak with the instructor during the first two weeks of class.

This syllabus is subject to change with reasonable advance notice by the instructor.